专利文本技术挖掘研究进展综述*

胡正银[1,2]　方　曙[1]
[1](中国科学院成都文献情报中心　成都 610041)
[2](中国科学院大学　北京 100049)

摘要: 【**目的**】归纳基于文本专利技术挖掘通用流程, 提炼其中关键技术, 并对典型挖掘场景进行分析。【**文献范围**】以“专利挖掘、专利分析”等关键词在 Elsevier、Springer、CNKI 数据库进行检索, 并参考全球技术挖掘相关会议, 共阅读相关文献 105 篇, 实际参考文献 66 篇。【**方法**】梳理其关键技术专利知识表示的研究现状与发展趋势, 选取三类典型技术挖掘场景进行分析, 通过归纳总结、提炼出专利技术挖掘未来发展趋势与研究热点。【**结果**】专利知识表示的粒度与结构决定了专利技术挖掘的深度、广度与维度。基于 SAO 基础语义单元, 面向技术难题与解决方案的专利技术挖掘有望成为未来发展趋势与研究热点。【**局限**】本研究仅探讨现有文本挖掘、统计分析、自然语言处理技术在专利技术挖掘中的应用情况, 对这些技术本身的发展趋势关注不足。【**结论**】本研究有助于全面了解专利技术挖掘的概貌、涉及的关键技术及主要应用场景。

关键词: 专利技术挖掘　语义知识表示　主题聚类　专利分类　技术演化
分类号: G353.1

1　引　言

技术挖掘是 21 世纪初, 美国学者 Porter 等提出的基于历史科技文献分析当前和未来技术发展现状与趋势的理论与方法, 研究范畴包括: 技术监测、技术竞争情报、技术趋势预测、技术路线图、技术评估、技术前瞻、技术流程管理、科技指标等[1]。技术挖掘的载体是已发表的科技文献, 包括: 学术论文、专利文献等。挖掘对象除了科技文献元数据字段, 如: 题名、作者、专利权人等外, 还深入到文献内容层面, 如: 文摘、全文、专利权利要求等。技术挖掘方法除基础统计分析外, 还包括: 知识抽取、文本挖掘、语义分析、数据可视化等。

与其他科技文献相比, 专利文献在文本表示方面具有格式规范、用语严谨等特点[2], 更容易表示成结构化语义模型; 在技术内容方面具有系统详尽、分类科学等特点[2], 更适合进行深度技术挖掘。因此, 专利文献成为技术挖掘使用最多的信息源。

专利技术挖掘内涵丰富, 从宏观层面上看, 可应用于未来技术趋势预测、技术前瞻研究等; 从中观层面上看, 可帮助研发机构进行技术监测、技术竞争情报分析等; 在微观层面上看, 可通过分析技术创新的基本元素、流程及方法, 作为改善和发明其他专利的基础, 为具体技术研发提供知识服务[1,3]。

2　专利文本技术挖掘

根据挖掘对象不同, 专利技术挖掘可分为: 专利元数据字段挖掘与专利文本挖掘两种。前者无论是方法论, 还是技术手段上都比较成熟; 而后者则是研究重点。

参考 Porter 等提出的通用技术挖掘流程与框架[1,4], 结合专利分析实际场景与流程, 归纳出专利文本技术挖掘流程如下: 确定待挖掘专利数据集、专利文本知识表示、分析专利技术挖掘场景、技术挖掘结果评估与修订。具体流程如图 1 所示, 其核心部分是: 专利文本知识表示与技术挖掘场景分析。

收稿日期: 2013-12-03
收修改稿日期: 2014-03-04
*本文系中国科学院西部之光项目“基于本体的专利文献技术挖掘系统研究与实践”的研究成果之一。

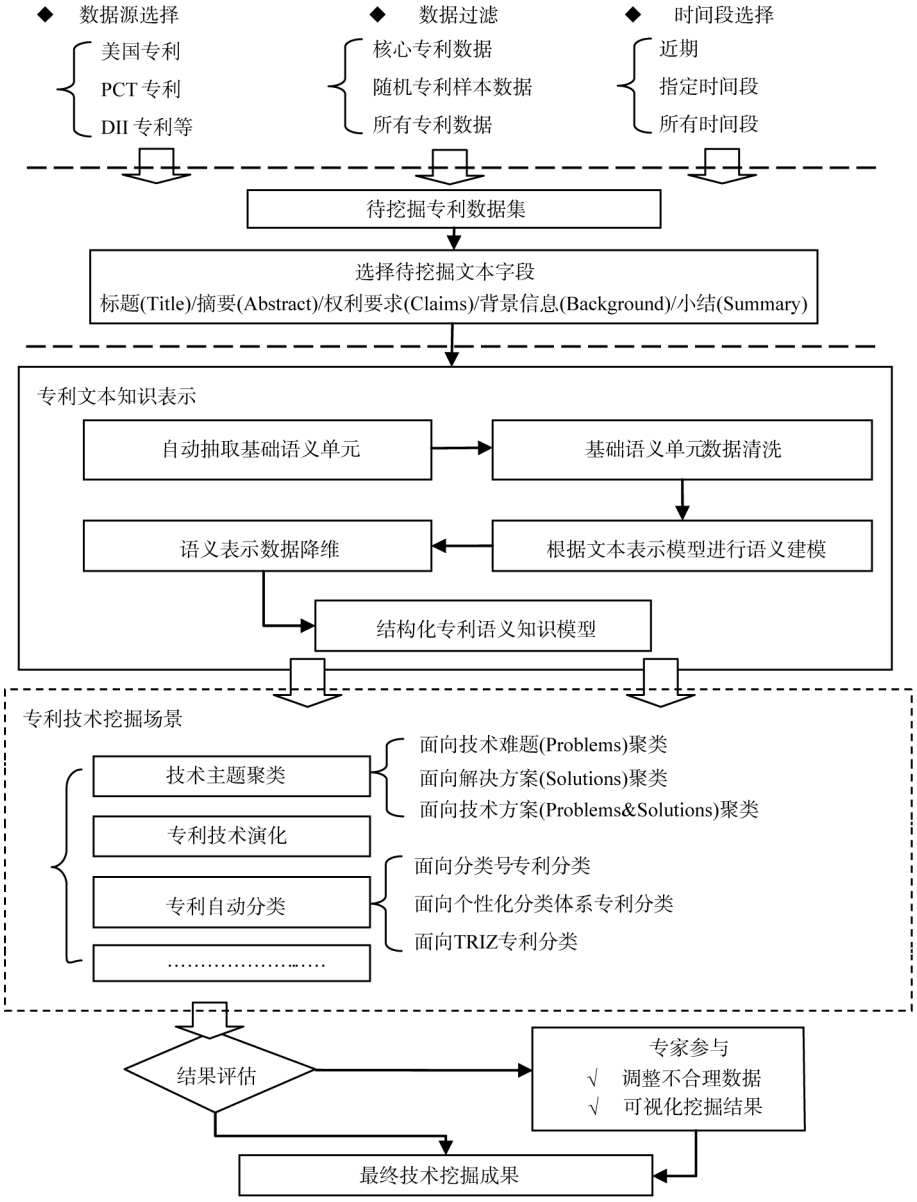

图 1　专利文本技术挖掘流程图

2.1　专利文本知识表示

专利文本知识表示是利用语义或文本表示技术, 从非结构化专利文本中抽取出能体现专利技术信息的结构化知识, 构建结构化语义模型。它是技术挖掘的基础与前提。根据语义建模方法不同, 现有研究可分为: 基于本体工程、基于向量空间模型、包含语义信息的向量空间专利知识表示三大类。

(1) 基于本体工程专利知识表示

该方法通过广泛专家咨询, 构建专利本体来进行知识表示。专利本体中不仅含有概念, 而且含有关系, 能充分、全面描述专利信息。它的典型研究项目是 PATExpert。 PATExpert 是欧盟 FP6 资助的一个项目, 目标是通过构建全面规范的本体来对专利进行全方位语义表示, 以支持语义层面上的专利分析与利用[5]。该项目核心专利本体包括: 专利元数据本体(Patent Metadata Ontology)、专利结构本体(Patent Structure Ontology)、专利内容本体(Patent Content Ontology)与专利图例本体(Patent Drawings Ontology)[6, 7]。PATExpert 定义的核心专利本体已成为其他类似研究的事实参考标准。此外, 还有一些针对特定应用的专

利本体研究, 如: IBM 印度研究院利用本体对生物医药类专利建模, 实现基于生物医药术语与关系的语义检索[8]; Ghoula 等对专利进行基于本体的语义标注, 为专利知识检索与推理提供数据支持[9]; 姜彩红等利用本体对中文专利进行知识抽取, 为构建专利知识库提供基础语料[10]。

该方法涉及的关键技术是本体映射与合并, 即: 如何准确、自动地将特定专利映射到标准化专利本体中。其优点是可多层次, 多角度地揭示专利语义信息。但缺点也显而易见: 构建专利本体需要大量专家参与, 费时费力, 不能满足敏捷场景下技术挖掘需求。

(2) 基于向量空间模型专利知识表示

向量空间模型(Vector Space Model, VSM)是一种经典的文本表示模型, 它凭借简单易懂的表示过程, 众多成熟的算法工具, 在文本表示中得到了广泛的应用。VSM 核心思想是将各种信息对象作为向量空间元素进行建模, 如: 文档、技术、概念、术语、检索等都表示成向量空间的向量。最简单的 VSM 是将专利文档直接表示成关键词权重向量。这种方法借鉴已有文本表示技术, 成熟易用, 在专利技术挖掘中得到广泛的应用, 如: Yoon 等研究以关键词向量为基础, 绘制专利技术关联图[11]。Lee 等用关键词向量表示产品资料、科技论文和专利文献, 通过分析关键词在三种文献中的分布, 制作基于关键词的技术路线图[12]。Kim 等研究了基于关键词向量的专利聚类及可视化方法, 用于新兴技术的预测[13]。

该方法涉及的关键技术是关键词向量降维, 即: 如何在准确完整地反映专利技术信息的前提下, 减少向量空间维度。其优点是简单易用。但由于关键词通常是名词或名词词组, 相互之间缺乏语义关系, 难以揭示专利深层次技术信息, 不能适用于微观层面的专利技术挖掘。

(3) 包含语义信息的向量空间专利知识表示

鉴于 VSM 专利知识表示的不足, 研究人员探讨利用三元组(Subject-Action-Object, SAO)来表示专利。它本质上依旧是一种 VSM, 但专利不再简单表示成关键词向量, 而是表示成内部具有一定语义关系的 SAO 向量。

早期, 基于 SAO 语义知识表示技术出于生物信息学研究的需要。生物学家希望能从大量的生物医学文献中挖掘出蛋白质与基因间的相互作用, 从而发现基因的特定功能[14-19]。鉴于 SAO 语义表示技术在生物信息学领域的巨大成功, 美国国家医学图书馆整合相关资源, 启动了面向生物医学领域的语义知识表示项目(Semantic Knowledge Representation, SKR)。SKR 以一体化医学语言系统(UMLS)为基础, 结合专家知识库系统, 从海量生物医学文献中自动抽取 SAO 结构, 实现生物医学文献的语义表示[20]。该方法很快被移植到专利上来, 如: Gaetano 等利用自然语言处理技术从专利中抽取 SAO 结构用于技术功效分析, 以此为基础开发了 PAT-Analyzer 系统[21-23]。Invent Machine 公司的语义 TRIZ 软件 GoldFire 中集成了 SAO 抽取功能, 并能自动将其分为 4 种类型[24, 25]: 组成(Made up of)、所属(Part of)、功效(Function)、相互作用(Interaction)。

研究人员结合专利文本特征, 进行了进一步研究, 如: Sheremetyeva 根据美国专利权利要求语言表述特点, 利用谓语词典, 将具有同义 Action 的 SAO 进行归并, 有效降低了 SAO 维度[26]。Yang 等将专利权利要求合并成效用(Utility)、功能(Functionality)、拥有(Contain)等 8 种语义类型, 大大降低了 SAO 维度[27]。在此基础上, 他们利用依存树将专利权利要求表示成基于 SAO 的语义概念图, 建立起 SAO 之间的语义关系[28]。Choi 等基于相似性对 SAO 进行分类, 构建技术分类树[29]。Choi 等在韩国国家自然科学基金的资助下进一步研究基于 WordNet 及专利功效分类的 SAO 本体构建, 实现了面向功效的专利检索[30]。Kim 等参照 TRIZ 中技术矛盾与发明原则将专利技术表示为技术难题与解决方案(Problems & Solutions, P&S), 并从词法–句法角度研究 P&S 表达模式[31]。Hu 等利用主题模型(Topic Model)进行 SAO 降维, 在此基础上实现了面向 P&S 的专利知识表示[32]。

该方法涉及的关键技术是 SAO 降维与知识重新表示。其优点是不需要像本体工程那样设计复杂语义知识模型, 语义映射较本体简单快捷; 相对关键词向量, 具有较丰富的语义表示能力。但现有研究大多没有考虑专利特有技术特征, 难以对 SAO 内部元素间语义关系进行深入挖掘。

(4) 小 结

总之, 基于本体工程专利知识表示需要大量专家参与, 费时费力; 基于 VSM 专利知识表示则因关键词

之间缺乏语义关系，难以应用于深层次技术挖掘。包含语义信息的向量空间专利知识表示则兼具两者优点，成为研究与应用的主流与热点。三类专利知识表示模型比较如表 1 所示:

表 1　三类专利知识表示模型比较

比较内容	基于本体工程专利知识表示	基于向量空间模型专利知识表示	含语义信息的向量空间专利知识表示
表现形式	专利本体	关键词向量	SAO 语义向量
模型复杂度	复杂度由专利本体 Schema 决定	简单	一般
语义表现力	丰富	贫乏	一般
研究现状	公开报道少，多为面向具体知识库应用	目前较少	研究热点，研究应用成果较多
研究热点	基于特定目的构建个性化专利本体	中观、宏观层面技术路线图分析	挖掘 SAO 内部元素间语义关系
关键技术	本体映射与合并	关键词向量降维	SAO 语义向量降维与知识重新表示
优点	多层次，多角度深入揭示专利语义信息	简单、成熟、易用	有一定语义表示能力、自动构建技术成熟
缺点	需要大量专家参与，费时费力	缺乏语义关系，难以揭示深层次信息	没有考虑专利特有技术特征
适用技术挖掘层次	微观层面上深入全面的技术挖掘	中观、宏观层面上基于概念的技术挖掘	微观、中观层面上具体技术内容挖掘

2.2　专利技术挖掘场景

专利技术挖掘范畴很广，本文选取技术主题聚类、专利自动分类与专利技术演化三类典型场景作为代表，分析其研究进展。

(1) 技术主题聚类

专利技术挖掘中，很少对专利本身进行聚类，多是对包含的技术主题进行聚类。根据聚类目的不同，可分为揭示技术主题分布、基于技术主题聚类的应用、研究技术主题之间的关系三类。

揭示技术主题分布是专利技术聚类最根本，也是最常见的应用。通常采用数据挖掘领域的一些成熟算法，如: K-means 聚类、层次累积聚类、多维尺度分析与自组织映射等对专利技术主题进行聚类。Tseng 等对专利分析中主题聚类进行了系统研究，归纳出流程如下: 关键词的选取、权重计算、相似度计算、聚类算法的选择、多步骤聚类、生成聚类簇标签、对聚类结果进一步分组等[33]。

基于技术主题聚类的应用则是将技术主题聚类作为其他应用的中间结果，如: Kang 等利用技术主题聚类结果来提升专利检索性能[34]。

研究技术主题之间的关系属于对聚类结果的进一步挖掘与分析，是目前专利聚类研究的重点。如: Kim 等在关键词聚类的基础上，计算关键词在不同聚类簇之间的数量分布，结合专利申请时间，分析出代表新兴技术的关键词[13]; Wang 等分析连接不同聚类簇特征关键词，将其作为技术过渡特征词予以研究[35]; Yoon 等结合多维尺度分析与离群点探测寻找独特的专利，认为这些专利有可能反映最新技术趋势[36]。

目前研究大多用关键词来表示技术主题，研究重点集中在技术主题术语收敛(Term Clumping)与技术主题间相似度计算。前者可使发散的关键词术语按照指定的规则收敛，为聚类提供高质量的基础数据; 后者则为复杂聚类及聚类后处理提供数据支持。

(2) 专利自动分类

专利分类既是组织管理专利的一种手段，也是技术挖掘的重要应用领域。根据分类体系不同，目前研究分为三类: 基于分类号自动分类、基于个性化分类体系自动分类、面向 TRIZ 自动分类。

分类号是专利领域最权威，应用最广泛的分类体系。目前，相关研究集中在如何建立更完善、统一的分类号体系，如: 欧洲专利局和美国专利商标局在 2010 年达成协议，在国际专利分类系统(International Patent Classification, IPC)基础上共同合作，创建并实施统一的联合专利分类系统(Cooperative Patent Classification, CPC)[37]。针对专利分类号系统层次结构复杂的特点，优化与改进分类算法也是一个关注热点。如: Fall 等基于 IPC 分类体系，比较朴素贝叶斯、支持向量机与 K-NN 三种分类算法的效果，通过优化训练集来提高一个专利分配多个分类号的准确率[38]。刘玉琴等基于 IPC 层次结构，针对不同层次类别建立

特征向量, 从而在各层次上实现专利自动分类[39]。Krier 等基于欧洲专利分类系统进行专利自动分类, 以便将相关专利申请分配给技术背景接近的审查员[40]。

专利分类号系统对技术描述过于宽泛, 难以满足特定目的技术挖掘需求。因此, 基于个性化分类体系的专利自动分类成为研究重点, 如: Falasco 基于美国专利分类系统, 实现根据产品功能与效果将专利进行再次分类[41]。Teichert 等研究了基于专利的 5 个功能类别进行专利自动分类的方法[42]。Lai 等在 IPC 基础上, 提出了应用于企业研发技术定位的专利分类方法[43]。Hu 等结合术语收敛与主题模型, 研究自动构建个性化的专利技术知识组织体系[44]。郭炜强等将 IPC 作为领域知识, 从专利文本中抽取特征关键词进行专利自动分类研究[45]。Kim 等对专利文档知识结构进行了重新组织, 将其分为技术领域(Technogical Field)、目的(Purpose)、方法(Method)、权利要求(Claim)、解释(Explanation)、实例(Example)等 6 个部分, 并在此基础上对专利进行分类[46]。

面向 TRIZ 自动分类则是依据 TRIZ 中发明原理或技术难题进行分类, 目的是帮助用户寻找利用了相似的发明原理或者解决了相似技术难题的专利。这些专利在技术领域上可能相差很远, 分布在不同的分类号体系中[47]。Loh 等对 TRIZ 40 条发明原理进行专利自动分类, 包括: 选取其中 6 条发明原理, 利用多种分类算法进行的小规模专利自动分类[48]; 分析 40 条发明原理之间的相似性, 对其进行重新分组[49]; 通过分析发明原理的专利句法信息, 归纳出相应的句法与语法模式, 基于关联规则进行面向发明原理的分类[50]。国内近两年也有类似的研究, 如: 梁艳红等进一步将发明原理归纳为显性发明原理与隐性发明原理两大类, 并实现了面向显性发明原理的专利自动分类[51]。翟继强等结合中文自然语言处理技术, 实现中文专利面向发明原理的自动分类[52]。

分类体系是自动分类的基础与前提。目前, 专利自动分类研究重点是构建个性化分类体系。面向个性化分类体系的专利分类中, 数据分布往往不平衡, 如果直接利用传统分类器进行分类, 准确率普遍较低[32]。非平衡数据集分类问题是专利自动分类中技术难点与关键点。

(3) 专利技术演化

技术演化分析技术主题产生、发展、突破创新、转移和变化乃至湮灭的过程, 是专利技术挖掘的重要内容之一。技术演化研究方法包括: 引文分析法、文本挖掘法、TRIZ 演化模型法、德尔菲法等。其中文本挖掘法、TRIZ 演化模型法常用于专利技术演化分析。

根据技术表示粒度不同, 可将文本挖掘法分为基于关键词与 SAO 两种。基于关键词的技术演化分析主要集中在中观或宏观层面, 如: Yoon 等基于专利关键词共词分析和形态学分析, 绘制移动电话领域技术演化路线图 [53,54]。Lee 等通过构建专利关键词演化地图, 发现新的技术机会[55]。方曙等研究关键词聚类簇分布随时间变化关系, 以此为基础发现新兴技术与基础技术[56]。

基于 SAO 分析专利技术演化是近几年出现的研究热点, 常与 TRIZ 技术系统 8 大进化法则[57]、9 窗口演化模型[58]结合起来进行微观层面技术演化分析, 如: Yoon 等系统结合 SAO 与离散点探测分析技术演化趋势, 从而识别新兴技术集群[36,59,60]。Park 等将 SAO 与 TRIZ 进化法则结合起来, 分析技术演化进程, 筛选领域重要专利[61,62]。Zhang 等基于 GoldFire 提供的语义 TRIZ 数据, 结合术语收敛与技术路线图, 发现与定位潜在新兴技术[63,64]。

基于关键词的技术演化分析相对成熟。基于 SAO 的技术演化分析是目前研究重点与热点, 涉及的关键技术包括: 基于 SAO 对技术进行语义表示、构建演化模型。

(4) 小 结

目前, 技术主题聚类基础是关键词; 专利自动分类基础是分类号、关键词; 专利技术演化分析的基础则是关键词与 SAO 基础语义单元。三种专利技术挖掘场景研究现状如表 2 所示:

表 2 三种专利技术挖掘场景研究现状

内容	专利技术主题聚类	专利自动分类	专利技术演化
基础	关键词	分类号、关键词	关键词、SAO
关键技术	术语收敛、技术主题相似度计算	非平衡数据集分类	基于 SAO 对技术进行语义表示
研究重点	对聚类结果进一步分析与应用	个性化分类体系分类、面向 TRIZ 分类	SAO 与 TRIZ 演化理论结合构建演化模型
不足之处	无法面向 P&S 聚类	无法面向 P&S 分类	无法自动挖掘 P&S 交替技术演化进程

3　结　语

专利文本技术挖掘的关键技术是专利文本知识表示，包含语义信息的向量空间专利知识表示是目前研究重点；典型场景包括：技术主题聚类、专利自动分类、专利技术演化等。专利知识表示的粒度与结构决定了各场景技术挖掘的深度、广度与维度。P&S 是专利技术的核心概念，专利知识表示能否准确揭示 P&S 信息，对专利技术挖掘成果优劣至关重要。现有研究总结如表 3 所示：

表 3　专利文本技术挖掘总结

关键技术	专利文本知识表示
典型场景	技术主题聚类、专利自动分类、专利技术演化
存在问题	①研究多集中在关键词或 SAO 层面，对专利特有 P&S 信息挖掘较浅； ②VSM 处理较大数据量级专利时，特征维度过大，效率低下。
改进方向	①面向 P&S 专利知识表示，即：基于 SAO，参照 TRIZ 技术矛盾与发明原则理论[47]，利用 Topic Model 定向生成 P&S 信息； ②概率模型与机器学习相结合进行较大数据量级专利文本技术挖掘。

总之，现有研究大都基于向量空间模型，在关键词或 SAO 层面对数据量较小专利集进行技术挖掘。Topic Model 是一系列基于概率模型，旨在发现大规模文档中隐含主题结构方法的统称[65]。与向量空间模型相比，Topic Model 在复杂知识表示、处理大数据量文本方面具有优势，如：潜在狄利克雷分配模型(Latent Dirichlet Allocation, LDA)可将海量科技文献表示成一系列主题的概率分布，主题表示成一系列关键词的概率分布[66]。相对于直接将文献表示成关键词或 SAO 向量，LDA 生成了新的技术特征(主题或 P&S)，更能揭示文献深层次知识结构。基于概率主题模型，面向 P&S 的专利技术挖掘将能深化专利技术挖掘程度，拓展专利技术挖掘范围，更新专利技术挖掘视角，有望成为未来的发展趋势与研究热点。

参考文献：

[1] Porter A L, Cunningham S W. Tech Mining: Exploiting New Technologies for Competitive Advantage[M]. Hoboken, New Jersey: John Wiley & Sons, Inc., 2005: 17-26.

[2] 王朝晖. 专利文献的特点及其利用[J]. 现代情报, 2008, 28 (9): 151-152, 156. (Wang Zhaohui. Characteristics and Utilization of Patent Documentation[J]. Modern Information, 2008, 28(9): 151-152, 156.)

[3] 吕详惠，仇宝艳，乔鸿. 基于本体的专利知识发现体系研究[J]. 计算机与信息技术, 2008 (7): 43-46. (Lv Xianghui, Qiu Baoyan, Qiao Hong. Study to Patent Knowledge Discovery Based on Ontology[J]. Computer and Information Technology, 2008(7): 43-46.)

[4] Porter A L, Zhang Y. Text Clumping for Technical Intelligence[EB/OL]. [2013-11-20]. http://www.intechopen.com/books/theory-and-applications-for-advanced-text-mining/text-clumping-for-technical-intelligence.

[5] Giereth M, Stäbler A, Brügmann S, et al. Application of Semantic Technologies for Representing Patent Metadata[C]. In: Proceeding of the 1st International AST Workshop, Informatik 2006, Dresden, Germany.2006.

[6] Wanner L. Advanced Patent Document Processing Techniques[EB/OL]. [2013-11-04]. ftp://ftp.cordis.europa.eu/pub/ist/docs/kct/patexpert-annualrep07_en.pdf.

[7] Wanner L, Baeza-Yates R, Brügmann S, et al. Towards Content-Oriented Patent Document Processing[J]. World Patent Information, 2008, 30(1): 21-33.

[8] Mukherjea S, Bamba B, Kankar P. Information Retrieval and Knowledge Discovery Utilizing a Biomedical Patent Semantic Web[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(8): 1099-1110.

[9] Ghoula N, Khelif K, Dieng-Kuntz R. Supporting Patent Mining by Using Ontology-based Semantic Annotations[C]. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence. IEEE, 2007: 435-438.

[10] 姜彩红，乔晓东，朱礼军. 基于本体的专利摘要知识抽取[J]. 现代图书情报技术, 2009(2): 23-28. (Jiang Caihong, Qiao Xiaodong, Zhu Lijun. Ontology-based Patent Abstracts Knowledge Extraction[J]. New Technology of Library and Information Service, 2009(2): 23-28.)

[11] Yoon B, Park Y. A Text-Mining-based Patent Network: Analytical Tool for High-Technology Trend [J]. The Journal of High Technology Management Research, 2004, 15(1): 37-50.

[12] Lee S, Lee S, Seol H, et al. Using Patent Information for Designing New Product and Technology: Keyword Based Technology Roadmapping[J]. R&D Management, 2008, 38(2): 169-188.

[13] Kim Y G, Suh J H, Park S C. Visualization of Patent Analysis

for Emerging Technology[J]. Expert Systems with Applications, 2008, 34(3): 1804-1812.

[14] Sekimizu T, Park H S, Tsujii J. Identifying the Interaction Between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts [C]. In: Proceedings of the 9th Workshop on Genome Informatics (GIW' 98), Tokyo, Japan.1998, 9: 62-71.

[15] Blaschke C, Andrade M A, Ouzounis C, et al. Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions[C]. In: Proceedings of the 7th International Conference on Intelligent System for Molecular Biology. The AAAI Press, 1999: 60-67.

[16] Thomas J, Milward D, Ouzounis C, et al. Automatic Extraction of Protein Interactions from Scientific Abstracts[C]. In: Proceedings of Pacific Symposium on Biocomputing. 2000: 541-552.

[17] Ono T, Hishigaki H, Tanigami A, et al. Automated Extraction of Information on Protein-Protein Interactions from the Biological Literature[J]. Bioinformatics, 2001, 17(2): 155-161.

[18] de Bruijin B, Martin J. Getting to the (C) ore of Knowledge: Mining Biomedical Literature[J]. International Journal of Medical Informatics, 2002, 67(1-3): 7-18.

[19] Jensen L J, Saric J, Bork P. Literature Mining for the Biologist: From Information Retrieval to Biological Discovery[J]. Nature Reviews Genetics, 2006, 7(2): 119-129.

[20] Aronson A R, Rindflesch T C. Semantic Knowledge Representation Project[EB/OL]. [2013-03-09]. http://skr.nlm.nih.gov/papers/references/BoSC98_rpt.pdf.

[21] Cascini G. System and Method for Performing Functional Analyses Making Use of a Plurality of Inputs: U.S. Patent Application US20050210382, European Patent Office EP1351156A1, International Publication Number WO20030 77154A3[P]. 2003-09-18.

[22] Cascini G, Rissone P. PAT-Analyzer: A Tool to Speed-up Patent Analyses with a TRIZ Perspective[C]. In: Proceedings of the ETRIA World Conference: TRIZ Future 2003, Aachen, Germany.2003.

[23] Cascini G, Fantechi A, Spinicci E. Natural Language Processing of Patents and Technical Documentation[C]. In: Proceedings of the 6th International Workshop on Document Analysis System. Berlin, Heidelberg: Springer-Verlag, 2004: 508-520.

[24] Verbitsky M. Semantic TRIZ[EB/OL]. [2012-09-10]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.115.1907&rep=rep1&type=pdf.

[25] IHS Inc. Optimize Decision-Making Across the Product Lifecycle[EB/OL]. [2012-12-02]. http://inventionmachine.com/Portals/56687/docs/OptimizingDecisionMakingAcrossthProductLifecycle_WhitePaper_InventionMachine.pdf.

[26] Sheremetyeva S. Natural Language Analysis of Patent Claims[C]. In: Proceedings of the ACL-2003 Workshop on Patent Corpus Processing.2003: 66-73.

[27] Yang S Y, Lin S Y, Lin S N, et al. Automatic Extraction of Semantic Relations from Patent Claims[J]. International Journal of Electronic Business Management, 2008, 6(1): 45-54.

[28] Yang S Y, Soo V W. Extract Conceptual Graphs from Plain Texts in Patent Claims[J]. Engineering Applications of Artificial Intelligence, 2012, 25(4): 874-887.

[29] Choi S, Park H, Kang D, et al. An SAO-based Text Mining Approach to Building a Technology Tree for Technology Planning[J]. Expert Systems with Applications, 2012, 39(13): 11443-11455.

[30] Choi S, Kang D, Lim J. et al. A Fact-oriented Ontological Approach to SAO-based Function Modeling of Patents for Implementing Function-based Technology Database[J]. Expert Systems with Applications, 2012, 39(10): 9129-9140.

[31] Kim Y, Tian Y, Jeong Y, et al. Automatic Discovery of Technology Trends from Patent Text[C]. In: Proceedings of the 2009 ACM Symposium on Applied Computing. New York, NY, USA: ACM, 2009.

[32] Hu Z Y, Fang S, Liang T. Automatic Patent Classification Oriented to Problems & Solutions[C]. In: Proceedings of Conference on Artificial Intelligence and Data Mining (AIDM 2013), Sanya, China.2013.

[33] Tseng Y H, Lin C J, Lin Y I. Text Mining Techniques for Patent Analysis[J]. Information Processing & Management, 2007, 43(5): 1216-1247.

[34] Kang I S, Na S H, Kim J. Cluster-based Patent Retrieval[J]. Information Processing & Management, 2007, 43(5): 1173-1182.

[35] Wang M Y, Chang D S, Kao C H. Identifying Technology Trends for R&D Planning Using TRIZ and Text Mining[J]. R&D Management, 2010, 40(5): 491-509.

[36] Yoon J, Kim K. Detecting Signals of New Technological Opportunities Using Semantic Patent Analysis and Outlier Detection[J]. Scientometrics, 2012, 90(2): 445-461.

[37] Kisliuk B. Introduction to the Cooperative Patent Classification(CPC)[EB/OL]. [2013-10-10]. http://www.uspto.gov/about/advisory/ppac/120927-09a-international_cpc.pdf.

[38] Fall C J, Törcsvári A, Benzineb K, et al. Automated

Categorization in the International Patent Classification[J]. ACM SIGIR Forum, 2003, 37(1): 10-25.

[39] 刘玉琴, 桂婕, 朱东华. 基于 IPC 知识结构的专利自动分类方法[J]. 计算机工程, 2008, 34(3): 207-209. (Liu Yuqin, Gui Jie, Zhu Donghua. Automated Categorization of Patent Based on IPC Knowledge Construction[J]. Computer Engineering, 2008, 34(3): 207-209.)

[40] Krier M, Zaccà F. Automatic Categorisation Applications at the European Patent Office[J]. World Patent Information, 2002, 24(3): 187-196.

[41] Falasco L. Bases of the United States Patent Classification[J]. World Patent Information, 2002, 24(1): 31-33.

[42] Teichert T, Mittermayer M A. Text Mining for Technology Monitoring[C]. In: Proceedings of 2002 IEEE International Engineering Management Conference(IEMC' 02). IEEE, 2002: 596-601.

[43] Lai K K, Wu S J. Using the Patent Co-Citation Approach to Establish a New Patent Classification System[J]. Information Processing & Management, 2005, 41(2): 313-330.

[44] Hu Z Y, Fang S, Zhang X, et al. Empirical Study of Constructing Knowledge Organization System of Patent Documents Using Topic Model[C]. In: Proceedings of the 2nd Global TechMining Conference, Montreal, Canada.2012.

[45] 郭炜强, 戴天, 文贵华. 基于领域知识的专利自动分类[J].计算机工程, 2005, 34(23): 52-54. (Guo Weiqiang, Dai Tian, Wen Guihua. A Patent Classification Method Based on Domain Knowledge[J]. Computer Engineering, 2005, 34(23): 52-54.)

[46] Kim J H, Choi K S. Patent Document Categorization Based on Semantic Structural Information[J]. Information Processing & Management, 2007, 43(5): 1200-1215.

[47] Mazur G. Theory of Inventive Problem Solving (TRIZ) [EB/OL]. [2013-08-12]. http://www.mazur.net/triz/.

[48] Loh H T, He C, Shen L. Automatic Classification of Patent Documents for TRIZ Users[J]. World Patent Information, 2006, 28(1): 6-13.

[49] He C, Loh H T. Grouping of TRIZ Inventive Principles to Facilitate Automatic Patent Classification[J]. Expert Systems with Applications, 2008, 34(1): 788-795.

[50] He C, Loh H T. Pattern-oriented Associative Rule-based Patent Classification[J]. Expert Systems with Applications, 2010, 37(3): 2395-2404.

[51] 梁艳红, 檀润华, 马建红. 面向产品创新设计的专利文本分类研究[J]. 计算机集成制造系统, 2013, 19(2): 382-390. (Liang Yanhong, Tan Runhua, Ma Jianhong. Study on Patent Text Classification for Product Innovative Design[J]. Computer Integrated Manufacturing Systems, 2013, 19(2): 382-390.)

[52] 翟继强, 王克奇.依据 TRIZ 发明原理的中文专利自动分类[J]. 哈尔滨理工大学学报, 2013, 18(3): 1-5. (Zhai Jiqiang, Wang Keqi. Automatic Classification of Chinese Patents According to TRIZ Inventive Principles[J]. Journal of Harbin University of Science and Technology, 2013, 18(3): 1-5.)

[53] Yoon B, Park Y. Development of New Technology Forecasting Algorithm: Hybrid Approach for Morphology Analysis and Conjoint Analysis of Patent Information[J]. IEEE Transactions on Engineering Management, 2007, 54(3): 588-599.

[54] Yoon B, Phaal R, Probert D. Morphology Analysis for Technology Road Mapping: Application of Text Mining[J]. R&D Management, 2008, 38(1): 51-68.

[55] Lee S, Yoon B, Park Y. An Approach to Discovering New Technology Opportunities: Keyword-based Patent Map Approach[J]. Technovation, 2009, 29(6-7): 481-497.

[56] 方曙, 胡正银, 庞弘燊, 等. 基于专利文献的技术演化分析方法研究[J]. 图书情报工作, 2011, 55(22): 42-46. (Fang Shu, Hu Zhengyin, Pang Hongshen, et al. Study on the Method of Analyzing Technology Evolution Based on Patent Documents[J]. Library and Information Service, 2011, 55(22): 42-46.)

[57] Petrov V. The Laws of System Evolution[EB/OL]. [2013-08-12]. http://www.triz-journal.com/archives/2002/03/b/.

[58] Mann D. System Operator Tutorial: - 1) 9-Windows on the World[EB/OL]. [2013-08-12]. http://www.triz-journal.com/archives/2001/09/c/index.htm.

[59] Yoon J, Kim K. An Automated Method for Identifying TRIZ Evolution Trends from Patents[J]. Expert Systems with Applications, 2011, 38(12): 15540-15548.

[60] Yoon J, Kim K. Identifying Rapidly Evolving Technological Trends for R&D Planning Using SAO-based Semantic Patent Networks[J]. Scientometrics, 2011, 88(1): 213-228.

[61] Park H, Ree J J, Kim K. An SAO-based Approach to Patent Evaluation Using TRIZ Evolution Trends[C]. In: Proceedings of the 6th International Conference on Management of Innovation and Technology(ICMIT). IEEE, 2012.

[62] Park H, Ree J J, Kim K. Identification of Promising Patents for Technology Transfers Using TRIZ Evolution Trends[J]. Expert Systems with Applications, 2013, 40(2): 736-743.

[63] Zhang Y, Porter A L, Hu Z Y. An Inductive Method for "Term Clumping": A Case Study on Dye-Sensitized Solar Cells[C]. In: Proceedings of the International Conference on Innovative

Methods for Innovation Management and Policy, Beijing, China.2012.

[64] Zhang Y, Porter A L, Gomila J MV, et al. Discovering Emerging Technology Trends: With TRIZ and Technology Road Mapping[C]. In: Proceedings of the 2nd Global TechMining Conference, Montreal, Canada.2012.

[65] Blei D M. Probabilistic Topic Models[EB/OL]. [2013-06-12]. https: //www.cs.princeton.edu/~blei/kdd-tutorial.pdf.

[66] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

作者贡献声明：

胡正银：研究过程实施，进行具体文献调研、分析与论文撰写；
方曙：研究命题的提出、设计，论文修订。

(通讯作者：胡正银 E-mail: huzy@clas.ac.cn)

Review on Text-based Patent Technology Mining

Hu Zhengyin[1,2] Fang Shu[1]
[1](Chengdu Library, Chinese Academy of Sciences, Chengdu 610041, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: [**Objective**] This paper generalizes the framework of patent technology mining based on text, extracts the key techniques and analyzes some typical application scenarios. [**Coverage**] Chooses 105 papers from Elsevier, Springer, CNKI databases and Global TechMining Conference, and refers 66 papers at last. [**Methods**] Review semantic knowledge representation of patents, analyze the research progress of three typical technology mining scenarios and summarize the hot research topics of patent technology mining based on text. [**Results**] The result shows that the semantic knowledge representation of patents is very important to patent technology mining. And patent technology mining oriented to problems and solutions based on SAO units will be the hot research topics. [**Limitations**] Only focuse on the applications in patent technology mining of the techniques (e. g. Text Mining, Statistics and Natural Language Processing), but the development trendency of these techniques need to pay more attention. [**Conclusions**] This paper will facilitate to give an overview of patent technology mining, the key problems and the typical application scenarios.

Keywords: Patent technology mining Semantic knowledge representation Topic clustering Patent classification Technology evolution